PATTERNS OF POWER

The Exploit

✦

Human Cognitive Architecture and the Permanent Vulnerability to Elite Capture — Why the Belief That Makes a Hierarchy Cheap Rests on Trust Machinery No Mind at Scale Can Do Without; That a Person Can See Through the Formation He Was Given, and Many Do, but Cannot See Without Trusting Someone, So That Every Escape From One Authority Is Completed by Deference to the Next; and That the Vulnerability Is Therefore Not a Curable Ignorance, and Not a Life Sentence in One's Childhood Creed, but the Standing, Capturable Price of Knowing in a World Too Large to Verify Alone — a Revision of the Paper of the Same Name, Holding the Framework as Tested, Not Revealed

THOMAS J. BRYAN

A STUDY IN THE PATTERNS OF POWER

Preface

THIS PAPER ANSWERS THE QUESTION THE WHOLE SERIES HAS circled without quite naming: not *how* a small governing class holds a vast population through managed belief — the other papers trace that in case after case — but *why it works at all*. Why do human beings, who are not stupid, who have long and bitter experience of being deceived, who mock the credulity of others as a favourite sport, so reliably accept the story that happens to serve whoever governs them? The answer is not that people are fools. It is that the very machinery which makes a person a competent knower in a world too large to check for himself — the machinery of trust, of deference to those who seem to know, of belonging to the group that carries the knowledge — is the same machinery that a governing class learns to capture. The vulnerability is not a defect laid over an otherwise sound mind. It is the underside of the mind's greatest strength. That is what this paper calls the exploit.

There is an earlier paper of this same name, and this one replaces it, because the earlier one contained a load-bearing error that has to be corrected in the open. The old paper argued, rightly, that the cognitive architecture cannot be removed — that no revolution and no education will breed the trusting, deferring, belonging, sacralising animal out of the species. But from that true premise it slid to a false conclusion: that the *person*, therefore, cannot really change — that recognition is not repair, that seeing through one's formation leaves one still bound to it, that the childhood creed is a kind of life sentence. This is wrong, and the corpus has since shown it to be wrong. People do see through the formation they were given. They deconvert and convert, defect and are converted, walk out of the church and the party and the family creed and build another life on other foundations; the contents of what a human being trusts are among the most

variable and historically plastic things about us. The framework a person is handed *can* change. What cannot change is the architecture underneath it — and the architecture is not a particular belief but the necessity of belief-through-trust as such. The correction, stated once and plainly, is this: **the framework is revisable; the architecture is not; and because the architecture is the human dependence on trusted others for nearly everything one knows, every escape from one formation is completed by entry into another.** The man who stops trusting the collar does not stop trusting. He trusts the byline, the credential, the badge of the new movement, the voice in the feed that finally seems to be telling him the truth. He has changed which authority he cannot verify. He has not — because no one can — stopped depending on authorities he cannot verify.

This is the cognitive engine underneath the paper that precedes this one in the sequence, *The Altar They Could Not Empty*, which followed the twentieth century's great attempt to abolish managed belief by force and found that the emptied altar always refilled — with the old god revived or a new one the state was compelled to build. That paper showed the pattern at the scale of nations. This one shows why the pattern is written into the single human mind: you cannot empty the seat of trusted belief, because a mind that had emptied it could not know anything at all. There is a discipline to hold throughout. The paper never asks whether any belief installed through this machinery is *true* — it asks only how belief *functions* and how the function is captured; a thing can be sincerely believed, load-bearing, and true all at once, and the analysis is silent on the last. And it holds, as its own governing example, a fact it cannot escape: this very revision was made the only way any revision is ever made — not by climbing out of trust into some view from nowhere, but by re-grounding on a refreshed body of evidence its authors now defer to. The paper corrects itself by the mechanism it describes. That is not an embarrassment. It is the point.

A Note on Sources and Collaboration

The framework is thomas J. Bryan's — the floor of force with a managed belief laid over it to make the force cheap, formation setting the benefit dial early, the sword and the community holding the cost dial, and the seam through which a people's beliefs are kept. So is the insight this paper turns on, which arrived, as his insights characteristically do, through pattern rather than through theorem: a recognition, carried across many historical instances before it had a vocabulary, that deference attaches to the *mark* of an authority and transfers across anyone who bears it — that the faithful believe not a particular priest but the collar, and any wearer of it. The name *verification override*, for the selective standing-down of a person's ordinary scepticism before an authorised source, is his coinage, an attempt to give that recognised pattern a word. The deep-time and intellectual history, the verification and refresh of the scholarship, the structure and the composition are Claude's.

Because the load-bearing evidence here is modern and copyrighted — the cognitive science of trust, cultural evolution, sacred values, and obedience — it is paraphrased and credited, with any quotation kept short; the technical terms of art (*epistemic vigilance*, *credibility-enhancing displays*, *the illusion of explanatory depth*) are named because they are the coin of the field. Where a finding is contested — the precise reading of Milgram's experiments, the exact size of Dunbar's number, the reach of the dual-process model — the paper says so rather than borrowing a certainty the science has not earned. A thematic bibliography closes the paper. This version supersedes the earlier paper of the same name and corrects the error described above.

The Argument in Seven Parts and a Coda

1 — The Collar, Not the Man. The everyday instance that opens the case: deference attaches to the mark of office and transfers across anyone who wears it. Name the puzzle — belief holds hierarchies, and it works on the intelligent and the suspicious — and reject the lazy answer that people are dupes.

2 — The Animal That Survives by Trusting. The deep history. We are the species that outsources cognition to the group; our success is built on obligate social learning. The trust machinery is not a weakness bolted on — it is the adaptation that made us dominant, which is exactly why it cannot be removed.

3 — The Vigilant Mind. The correction to the gullibility story. People are not blank credulity; they run open vigilance, trusting on conditions — competence, benevolence, consensus, coherence, costly commitment. The cues vigilance reads are the attack surface.

4 — The Five Levers. The components of the exploit, re-grounded in the science: the deference heuristic, coalitional belonging, sacred values, narrative and intuition, and the timing of formation — and the verification override as the name for all five acting as one.

5 — The Mismatch and the Capture. In the small world the cues were honest. At civilizational scale the cue and the competence come apart, so the cues become forgeable; the governing class learns, by trial and error, to manufacture them — and to form the child before vigilance matures.

6 — Seeing Through Is Not Getting Out. The centre. The framework is revisable and people do revise it; but the architecture is the division of cognitive labour, and no one knows more than a sliver first-hand — so seeing through one authority only ever relocates trust to another. The permanence is in the dependence, not in the mind's rigidity.

7 — The Strongest Counter. That mass persuasion mostly fails and people are hard to fool — so perhaps there is no exploit, only ordinary, well-functioning trust. Stated at full strength, and answered: the exploit satisfies vigilance rather than defeating it, and dependence can be relocated but never exited.

Coda — The Crowded Seam. What follows. Since no one can be freed from trusting authorities, the only liberty available is in the structure of the authorities one must trust: plural, checkable, revisable, competing. Not the end of dependence — its contestability.

PART 1

The Collar, Not the Man

> *That the whole argument can be entered through a single ordinary observation, of the kind that furnishes belief across every ranked society: that in a devout Catholic family the word of a priest is taken as the word of the Church and of God, and taken so not because of anything known about the particular man — his character, his learning, his track record, none of which the believer has examined — but because he wears the collar, which certifies that he speaks for the institution; so that the deference is not to a person at all but to a mark, and transfers intact to any stranger who bears it; and that this everyday transfer of trust to a symbol, which stands in for a verification the believer will never perform, is the exploit in its simplest visible form, and the thing this paper must explain — not as the credulity of simple people, for it operates in the sophisticated exactly as in the simple, but as a feature of how any mind must work in a world it cannot check for itself*

Begin with something small enough to see whole. In a devout Catholic household a priest comes to the door, and what he says about God, about the sacraments, about how a life should be lived, is received as the word of the Church and, behind the Church, of God. Ask what the family knows about this particular man — his theology, his private conduct, whether his judgement is sound or his learning deep — and the answer, almost always, is nothing. They have not examined him. They do not need to, and it would not occur to them to try. He wears the collar. The collar certifies that he speaks not for himself but for the institution; and the deference the family extends is not to him but through him, to the thing the

collar marks. Send a different priest tomorrow, an utter stranger, and the same deference attaches at once, undiminished, because it was never about the first man in the first place. The trust is fixed to the mark, and the mark is transferable.

This is worth pausing on, because it is the exploit in its simplest visible form, and once seen here it can be seen everywhere. The believer has performed no verification. He could not: he has neither the training to assess the theology nor the access to assess the man, and even if he did, he has a life to live and cannot spend it auditing the credentials of everyone whose word he must take. So the collar does the work that verification would do if verification were possible — it stands in for the check the believer will never run. This is the mechanism the framework has elsewhere named the *verification override*: not a general gullibility, but the selective setting-aside of ordinary scepticism for a source the collar has marked as authorised. The same person who would question a stranger's claim about anything else, who haggles in the market and disbelieves the salesman and reads the neighbour's motives shrewdly, stands the scepticism down the moment the authorised mark appears. He is not less intelligent when he does this. He is doing what a mind in a world of specialised knowledge has to do: taking on trust, from a marked source, what he has no way to establish for himself.

And here is the first thing the old account got wrong and this one must get right. It is tempting to read the scene as the credulity of simple people — to suppose that the family defers to the collar because they are unlettered or unworldly, and that a more sophisticated person would not. But the transfer of trust to a mark is not a defect of the simple; it is universal, and it operates in the learned exactly as in the unlearned, only with different collars. The physicist takes on trust almost the entire edifice of biology, as the biologist takes the physics; the judge trusts the forensic examiner whose methods she cannot audit; the surgeon trusts the metallurgist who certified the steel;

every one of them, off duty, trusts the pilot, the pharmacist, the engineer who built the bridge they drive across without a thought. None of these is a fool. Each is doing, with a credential or a licence or a peer-reviewed journal in the place of a collar, precisely what the devout family does: extending trust to a mark that certifies an authority they cannot themselves verify. The puzzle the series has circled is why managed belief holds hierarchies so cheaply and so durably, and why it works on the suspicious and the clever and not only on the meek. The answer this paper will give is that it works on everyone because it is not aimed at a weakness some people have. It is aimed at the thing all knowing at scale is made of. To see why, one has to go back to the animal that had to build its knowledge this way or not survive at all.

PART 2

The Animal That Survives by Trusting

> *That the trust machinery is not a flaw the mind would be better without, but the single adaptation on which the human career was built — because the human strategy for survival was never individual cleverness, of which any one person has pitifully little, but the pooling and transmission of a body of knowledge no individual could reinvent or even fully understand, held across the group and handed down the generations; that this strategy requires each person to acquire most of what he knows by copying and believing others rather than by checking for himself, and to be biased toward copying the successful and the many; that the disposition to defer, to belong, and to take the group's knowledge on trust is therefore not a bug bolted onto reason but the enabling condition of the only kind of intelligence our species actually has, which is collective; and that a machinery this deep and this load-bearing cannot be removed without removing what makes us human — so that any project premised on freeing people from the need to trust is not reform but fantasy, aimed at an animal that has never existed*

To understand why the exploit cannot be patched, one has to abandon a flattering picture of the human mind — the picture of the lone rational agent who, given good information and freedom from manipulation, works things out for himself. That agent does not exist, and if he did he would not have survived. The actual human strategy, the thing that made a physically unimpressive primate the dominant animal on the planet, was never individual cleverness. Any single human, stripped of the knowledge his group hands him, is a hopeless creature: drop him naked

INTO A LANDSCAPE HIS ANCESTORS MASTERED AND HE WILL DIE of exposure or hunger within days, unable to reinvent from scratch the fire, the tools, the tracking lore, the knowledge of which plants heal and which kill that took his people a thousand generations to accumulate. What made humans formidable was not that each of us is smart but that we are, uniquely, an animal that pools cognition — that builds across generations a body of practical knowledge no individual could produce or even wholly comprehend, and transmits it by teaching and copying so that each new person inherits the accumulated competence of all who came before. The cultural evolutionist Joseph Henrich has made this case at book length: the secret of our success is not our brains considered one at a time, which are not up to the tasks our survival requires, but our capacity for cumulative culture, which makes the group smart in a way no member is.

This strategy has a price built into it, and the price is the exploit. If your survival depends on acquiring a vast body of knowledge that you cannot verify or reconstruct yourself, then you must acquire most of it by *trusting* — by copying what others do and believing what others tell you, on a scale that dwarfs anything you could ever check. A child cannot demand the evidence for why the tuber must be leached before eating, or the reasoning behind each step of making the poison for the arrows; the causal chains are too long and too opaque, and the child who insisted on verifying before believing would be dead before he finished. So the mind is built to take the culture on trust — and, because not every model is equally worth copying, to take it on *biased* trust, weighted toward particular others. Two of these biases matter most here, and both are well documented. One is the bias toward the many: other things equal, copy what most people around you do, because a practice common in a surviving population has usually been filtered by results. The other is the bias toward the successful and the prestigious: preferentially attend to and imitate those who display skill or command the deference of others, on the sound heuristic that people worth

copying are the ones others have already found worth copying. Henrich and Francisco Gil-White traced how this second bias produces the human phenomenon of prestige itself — status freely conferred on those with valued skill or knowledge, in exchange for the chance to learn from them — and showed that we infer whom to defer to from cues: the attention others pay them, the markers of their skill, the very deference they already attract.

Notice what this means. The disposition to defer to the marked and the prestigious, to conform to the group, to take on trust what one cannot check — the very dispositions the exploit will capture — are not weaknesses that crept into an otherwise rational mind. They are the enabling conditions of the only intelligence our species has ever had, which is collective and transmitted, not individual and self-made. The evidence that they run deep is everywhere once one looks. Human children, unlike other apes, *over-imitate*: shown a sequence of actions to get a reward, they faithfully copy even the steps that are visibly pointless, the ones a chimpanzee skips — because the child's overwhelming default is to assume the knowledgeable adult is doing something worth copying even where the child cannot see why. That looks irrational, and in the narrow case it is; but it is the local signature of the general strategy that made us, the strategy of trusting the accumulated competence of others over one's own meagre first-hand grasp. The mind that copies the pointless step is the same mind that copies the leaching of the tuber, and only the second matters, but you cannot have the second discernment without the first disposition, because at the point of copying the child cannot tell which is which. This is the deepest reason the exploit is permanent, and it is a better reason than the one the old paper gave. The old paper said the architecture cannot be removed because evolution is slow — because the Pleistocene mind is stuck in a Space Age world and cannot be re-bred in any timeframe that matters. That is true as far as it goes. But it undersells the point, because it frames the trust machinery as an obsolete liability we are merely stuck with. The truth is stronger and stranger: the

machinery is not obsolete at all. It is doing, right now, the thing it was always for — making us collectively knowledgeable by making each of us individually dependent — and a species that switched it off would not be a liberated humanity but no humanity at all, a scatter of naked animals who knew only what each could verify alone, which is almost nothing. You cannot free people from the need to trust. There is no one under there who could survive it.

PART 3

The Vigilant Mind

> *That the machinery of trust would have been suicidal if it were mere credulity, because a creature that believed whatever it was told would be the helpless prey of every liar, and communication itself would collapse once lying paid better than truth-telling; that natural selection therefore built, alongside the disposition to trust, a suite of filters — epistemic vigilance — by which the mind screens what it is told, weighing the source for competence and for honest intent, checking the message against what it already holds, and attending to whether the speaker acts on his own claims at a cost; that human beings are in consequence not gullible but conditionally trusting, open enough to learn from others and guarded enough not to be routinely had, which is why crude mass persuasion so reliably fails; and that this is not a refutation of the exploit but the key to it — because vigilance does not verify the world, it reads cues that stand in for verification, and a cue that can be read can be forged, so that the exploit works not by defeating the mind's vigilance but by satisfying it with manufactured signals of competence, consensus, and sincerity that the vigilance, functioning exactly as designed, waves through*

A creature built only to trust would not have lasted a season, because trust without a filter is prey. The moment communication becomes possible, so does deception, and if believing whatever one is told paid worse than disbelieving it, listeners would stop believing and the whole apparatus of shared knowledge — the thing that made trust worth having — would collapse. So the same evolutionary pressure that built the disposition to trust built, inseparably, the disposition to screen: a suite of mechanisms the cognitive scientist Dan Sperber and his

colleagues named *epistemic vigilance*. We do not, in fact, believe just anyone about just anything. We weigh the source on two axes — is this person *competent* to know, and does this person have any interest in *misleading* me — and we check the message for internal coherence and for fit with what we already take to be true, resisting what clashes with our existing picture. Sperber's sometime collaborator Hugo Mercier has pressed the point into its strongest form: humans run what he calls *open vigilance*, open enough to be moved by good reasons and reliable sources, vigilant enough that we are, by default, hard to move at all. On this account the received image of the credulous mass — the public helplessly hypnotised by demagogues and advertisers — is largely a myth. Mass persuasion, Mercier argues from the historical record, mostly fails; propaganda rarely converts anyone who was not already disposed to agree; the demagogue usually preaches to a choir he did not create. People are not sheep. They are wary judges of what they are told, and the wariness works.

This is the correction the old paper most needed, and it does not weaken the argument — it is the argument's key. The exploit is not, and must not be stated as, the claim that people are gullible. That claim is false, and a theory of power built on a false psychology will mislead wherever it is applied. People are vigilant. The exploit works anyway, and the reason it works is the one thing about vigilance that the reassuring story leaves out: vigilance does not, because it cannot, verify the world. It reads *cues*. When the mind judges a source competent, it is not auditing the source's knowledge — it is reading the marks of competence: the credential, the manner, the setting, the deference of others, the fluency of the performance. When it judges a source sincere, it is not seeing into the heart — it is reading the signs of sincerity, above all the sign that carries the most weight because it is the hardest to fake in the small world: whether the speaker acts on his own professed belief at a cost to himself. Henrich named this last cue the *credibility-enhancing display* — we trust the belief of those who pay for it,

who fast and tithe and risk and sacrifice in ways that would be mad if they did not truly hold what they profess — and it is why the martyr persuades where the pamphlet does not, and why every durable creed has surrounded its authorities with costly, visible, believed-in devotion. Vigilance runs on these cues because in the world that shaped it the cues were reliable: the man who displayed the marks of skill generally had the skill, the man who paid for his belief generally held it, the practice most people followed had generally been vetted by survival. The filters were calibrated to an environment in which the sign and the thing it signified travelled together.

And that is the whole opening. A cue that can be read can be forged. The exploit does not need to defeat the mind's vigilance — a frontal assault on vigilance is exactly the crude persuasion that reliably fails. It needs only to *satisfy* vigilance, to present it with manufactured versions of the very signals it is built to trust: the marks of competence without the competence, the appearance of consensus without the independent agreement, the costly display of a sincerity that serves the displayer, the prestige that is conferred and amplified until the deference of others becomes its own evidence. Against a forged cue, vigilance does not malfunction. It functions perfectly, and waves the forgery through, because a correctly working filter that has been fed a convincing counterfeit returns exactly the answer the counterfeit was built to produce. This is why the exploit works on the clever and the suspicious. Their vigilance is if anything sharper — but it is still vigilance, still cue-reading, still blind to everything the cues do not carry, and a sufficiently well-made counterfeit satisfies a sharp filter as readily as a dull one, sometimes more readily, because the clever are more confident that they cannot be fooled. The collar is a manufactured cue of authorised competence. The show trial's confession is a manufactured credibility-enhancing display. The manufactured consensus of a captured press is a forged signal of the many. In each case the mind is doing its job correctly on corrupted inputs. The exploit is the corruption of the inputs.

PART 4

The Five Levers

> *That the exploit is not one vulnerability but a cluster of five, each a cue-reading disposition that was reliable in the small world and is forgeable in the large one, and that the governing classes learned to work individually and, at their most effective, all at once in a mutually bracing system: the deference heuristic, by which we defer to the marks of legitimate authority; the coalitional instinct, by which we conform to and stand with the group whose belonging we cannot do without; the sacred-value lock, by which certain commitments are placed beyond calculation and made non-negotiable; the narrative and intuitive channel, by which a story structures the ground before argument can examine it; and the timing of formation, by which all of the above are installed in the child before the vigilance that would screen them has matured; and that the five acting together are what the framework calls the verification override — the standing-down of scepticism for the authorised source that results when a marked authority, backed by one's community, wrapped in the sacred, delivered as story, and installed before one could refuse it, presents its claim*

The exploit resolves into five levers, and it is worth taking them one at a time before seeing how they brace one another, because each is a distinct cue-reading disposition and each is forged in a distinct way. Together they compose the verification override.

The first is the **deference heuristic** — the disposition, examined already through the collar, to take the word of a marked, legitimate authority without independent check. It was reliable in the band, where the person who held authority had usually earned it in front of everyone over years, so that the mark and the merit were the same thing. Its most famous

demonstration is also its most misunderstood. In Stanley Milgram's obedience experiments the majority of ordinary people were led to administer what they believed were dangerous shocks to a stranger, at the instruction of a man in a lab coat, and this was long read — including in the earlier version of this paper — as proof that human beings blindly obey the symbol of authority, that the coat itself compels. But the careful modern re-examination of Milgram's own records, by Alexander Haslam and Stephen Reicher, tells a different and more exact story. The participants were not blindly obedient; they were agonised, resistant, sensitive throughout to the victim's plight. What moved them to continue was not the raw command — the prod that simply ordered them on, *you have no choice*, was the *least* effective and provoked refusal — but the prods that invoked the *value of the scientific enterprise* they had been led to believe they were serving. They complied to the extent that they identified with a legitimate authority pursuing a cause they had been persuaded was worthy. This is *engaged followership*, not blind obedience, and it corrects the lever precisely: the deference heuristic is not a hypnotic response to a symbol that switches off the mind. It is the vigilant mind's rational deference to what it takes to be a legitimate authority in a worthy cause — which is why the exploit's task is never to hypnotise but to *legitimate*, to manufacture the authority and the worthiness that earn the deference the mind is glad to give.

The second is the **coalitional instinct** — the disposition to conform to one's group and to stand with it, rooted in the plain ancestral fact that expulsion from the band was death, so that the machinery generating the fear of exclusion and the pull of solidarity was not weakness but survival. Solomon Asch's experiments showed how far it reaches: people will call a line the wrong length against the plain evidence of their eyes rather than stand alone against a unanimous group — though, tellingly for the vigilance point, most people resisted on most trials, and the conformity was a pull, not a compulsion. Coalitional psychology is flexibly addressed: the mind

readily encodes *whoever* it is taught to see as its people, which is why manufactured groups — the nation, the confession, the party, the unit — can capture a solidarity that evolved for genuine survival communities, and can make defection from the governing arrangement feel like the old mortal thing, self-expulsion from one's own.

The third is the **sacred-value lock** — the disposition to place certain commitments beyond calculation, to treat them as absolute and to feel moral revulsion at the mere suggestion of trading them for advantage. Philip Tetlock documented how these protected values resist the trade-off reasoning that governs ordinary preferences, and how the attempt to price them produces not compromise but outrage; Scott Atran and Jeremy Ginges have shown how sacred values, fused with group identity, produce the *devoted actor* who will sacrifice everything for a cause and by whom small, materially weak groups outlast far stronger ones. In the band this machinery made commitments credible — the ally who would never, at any price, betray you was worth more than one who would if the price were right. Captured, it is what converts a contingent governing claim into something felt as holy, so that challenges to the arrangement register not as arguments to be weighed but as desecrations to be repelled, and the person who reasons against the sacred value produces not reconsideration but a closing of ranks.

The fourth is the **narrative and intuitive channel** — the fact, which the study of moral judgement has established, that the mind reaches its convictions first by fast intuition and only afterward by slow reasoning, and that its native format for the important things is not the proposition but the story. Jonathan Haidt's work showed the intuition leading and the reasoning following, arriving to justify a verdict already reached; and Mercier and Sperber have argued that reason itself evolved less to find truth in solitude than to produce and evaluate arguments in the social group. Story exploits this by colonising the ground before argument can walk it: a

compelling narrative, absorbed in childhood and charged with emotion, installs the cast of the world — who the heroes are, what the founding sacrifice was, what the order is for — before the deliberative faculty is ever invited to assess a claim, so that when assessment finally comes it proceeds across a landscape the story has already shaped, and reaches the conclusions the landscape was built to yield.

The fifth is not a mechanism but a matter of **timing** — the sunk cost of formation. All four of the preceding levers are installed most durably when they are installed early, in the child, before the vigilance that would screen them has matured and while the longest runway remains for a whole identity to be built on top of the foundation. The proverbial Jesuit boast — give me the child until seven and I will show you the man — is a precise statement of the mechanism: not that children are easier to argue into things, for they are not, but that a belief laid down before one could refuse it, and then buried under decades of a life built upon it, is defended in adulthood not as a proposition one might update but as the ground of the self, whose revision would be a kind of self-demolition. The oath, the initiation, the uniform, the first communion — each adds a further public, identity-fusing layer, raising the cost of any later defection.

These are the five levers, and their real power is combinatorial. Singly, each can be resisted; a vigilant adult can doubt an authority, stand against a group, question a sacred claim, see through a story. But the governing classes learned — not by theory but by the long, ruthless trial and error of the craft of rule — to work all five at once, so that they brace each other: the authority is also the head of your community, and questioning it is disloyalty; the loyalty is also sacred, and disloyalty is sin; the sacred order arrives as the story of your people, absorbed before you could weigh it; and by the time you could weigh it your whole life was built on it. When a marked authority, backed by one's belonging, wrapped in the sacred, delivered as the story of one's people, and installed before one could refuse

it, presents a claim, the vigilance that would screen any one of these stands down before all five together. That standing-down is the verification override, and it is not stupidity and not cowardice. It is a vigilant mind, working correctly, meeting a set of cues each of which it was built to trust, all pointing the same way at once.

PART 5

The Mismatch and the Capture

> *That the levers were not vulnerabilities in the world that shaped them, because in the small band of known persons the cue and the thing it certified were bound together — the authority had been watched earning his standing, the group was a real community of shared fate, the sacred commitment was to actual kin, the story carried tested survival knowledge, and formation installed what the group had genuinely learned; that the agricultural revolution and the state broke that binding by exploding the scale, producing millions who would never meet their rulers or verify the institutions they were told to obey, so that for the first time the cue and the competence, the symbol and the merit, could come apart — and once they could come apart, they could be pulled apart deliberately; that the governing classes discovered, through millennia of what worked, how to manufacture each cue the vigilant mind was built to trust; and that this is why the exploit targets childhood and the shared environment rather than the reasoning of adults — because adult vigilance resists direct persuasion, as the sceptics rightly insist, and the way around a vigilance you cannot defeat is to reach the person before it matures and to counterfeit the signals it will later, correctly, trust*

THE LEVERS WERE NOT DEFECTS IN THE WORLD THAT BUILT them, because in that world the cue could be trusted — not by luck but because the cue and the thing it stood for were physically bound together and could not easily be prised apart. In a band of perhaps fifty known people, the authority you deferred to was the elder or the master

HUNTER WHOSE COMPETENCE YOU HAD PERSONALLY WATCHED accumulate across your whole life, so the mark of his standing and the merit behind it were one and the same, and deference to the mark was deference to demonstrated skill. The group whose belonging you could not do without was a genuine community of shared fate, in which conformity really did track collective survival. The sacred commitments were to actual kin with whom your life was actually interwoven. The stories carried real, tested knowledge of a real, shared world. And formation installed in the child what the group had genuinely learned. In that world the levers were not exploitable, because there was no gap between the signal and the substance for an exploiter to work in. The verification override was not an override of anything worth doing; it was the sensible economy of a mind that could take the marked source on trust because, there, the marked source was reliable.

The agricultural revolution and the state that grew from it broke the binding by exploding the scale, and the paper's other studies trace the consequences institution by institution. Suddenly there were societies of hundreds of thousands and then millions, governed by people the governed would never meet, organised through offices and credentials and ceremonies that no individual could trace to any demonstrated competence, holding together communities of strangers who shared a flag but not a life. For the first time the cue and the thing it certified could come apart: the man could wear the mark of an authority he had not earned, the crowd could be made to display a consensus none of them had independently reached, a costly ritual could be staged to manufacture a sincerity that served its stager, a story could be composed to install a cosmos that served its authors. And the moment the signal and the substance *could* be separated, they *could be separated on purpose* — which is the whole opening the governing classes moved into. They did not discover this through any theory of cognition; they discovered it the way the craft of rule discovers everything, by the brutal

winnowing of what worked, in which the priesthood that found the creation story produced more reliable labour than the garrison, and the dynasty that found the anointing produced more reliable loyalty than the wage, were rewarded with survival, and their competitors who missed it were replaced. What accumulated across the millennia was a practical mastery of the forgery of cues — how to manufacture the marks of legitimate authority, the appearance of the many, the displays of costly sincerity, the stories that install the order as the shape of the world — a mastery held not as written doctrine but as the transmitted know-how of ruling institutions, refined by people who were rewarded for getting it right.

This is also the answer to the strongest objection, which the next part will state at full strength but which must be met here in outline, because it bears directly on how the capture actually works. The sceptics are right that adult vigilance resists direct persuasion — that you cannot, as a rule, argue a wary grown person out of his interests or into a creed he has reason to doubt, and that the crude image of the propagandist reprogramming the passive masses is mostly false. But the exploit was never a frontal assault on adult vigilance, and reading it as one is the error that lets the objection seem fatal. The exploit goes around the vigilance it cannot defeat, by two routes the whole architecture makes available. The first is *timing*: it reaches the person in childhood, before the filters that would screen the installation have matured, and lets the sunk cost of a life built on the foundation do the rest — which is why formation systems everywhere target the young and why the fight over schooling and upbringing is never incidental to power. The second is *cue-capture*: for the adult whose vigilance is mature and cannot be stormed, the exploit does not try to storm it; it feeds it counterfeits — the manufactured authority, the manufactured consensus, the manufactured display of sincerity — so that the vigilance, doing its job correctly, certifies the counterfeit and delivers the deference the counterfeit was built to earn. Against direct persuasion the adult mind is well defended,

exactly as the sceptics say. Against a childhood it did not choose and an information environment whose cues have been captured, it is defended by nothing, because the defence it has is cue-reading, and the cues are what the exploit owns.

PART 6

Seeing Through Is Not Getting Out

> *That here the earlier paper made its one grave error, and here it must be corrected: it held that because the architecture cannot be removed, the person cannot really change — that recognition is not repair, that the one who sees through his formation remains its prisoner, that the childhood creed is a life sentence; and this is false, for people do see through the formation they were given, and change it, all the time, which is the actual foothold of every liberty the series prizes; that what cannot change is not the framework but the architecture beneath it, and the architecture is the division of cognitive labour — the plain fact, now well documented, that no person knows more than a sliver of what he believes at first hand, that we vastly overestimate even that sliver, and that the overwhelming mass of anyone's knowledge is borrowed on trust from a community he cannot audit; so that to stop trusting one authority is never to stop trusting, but only to transfer the trust to another, the vigilance re-engaging on new cues, the levers re-seating on new materials, the verification override re-arming for the new source; and that the exploit is therefore permanent not because the mind is rigid but because knowing at scale is ineliminably dependent, and dependence is the thing that can be captured — the seat of trusted belief cannot be emptied, only refilled, in the single mind exactly as in the nation*

Now the correction, which is the reason for the re-write. The earlier paper, having established the true and important point that the cognitive architecture cannot be bred or educated out of the species, drew from it a conclusion that does not follow and is not true: that the *person*, in consequence, cannot really change — that recognition is not

repair, that the one who comes to see the mechanism remains bound by it, that the formation installed in childhood is a kind of sentence one serves for life. Read against the record, this is simply false, and the falseness matters, because it quietly forecloses the one thing every liberty the series values depends upon. People do see through the formation they were given. They lose the faith they were raised in and find another, or none; they walk out of the party whose catechism they once recited; they turn against the national story they saluted as children; they defect from the family creed at great personal cost because they have come to hold something truer above it. The whole comparative record of the previous paper — the believers who kept the faith the atheist state tried to abolish, and equally the believers who abandoned faiths their states did nothing to touch — is a record of formation being revised, in both directions, constantly. And the science agrees: the contents of human psychology are among the most variable things about us, so much so that the traits once taken for the universal human mind turn out, on inspection, to be the peculiar and historically recent products of one unusual cultural lineage. The framework a person carries is not fixed. It is plastic, revisable, and often revised. To deny this is to deny liberty its actual mechanism, which is precisely the widening of the field until a person can see, and come by his own weighing to hold, something other than what he was handed.

What cannot change is not the framework but the architecture, and the whole force of the correction lies in seeing that these are different things. The architecture is not any particular belief; it is the *necessity of belief through trust* — and that necessity is the division of cognitive labour, which the modern science of the mind has documented with a precision the intuition resists. No person knows more than a narrow sliver of what he confidently believes. Ask people to explain the ordinary mechanisms they are sure they understand — how a zipper works, how a flush toilet works, how a bicycle stays up — and their confidence collapses the moment they

try, revealing an *illusion of explanatory depth*: we routinely mistake a vague familiarity for real understanding, and the mistake is largest exactly for the causal, mechanistic knowledge on which we act. And the sliver we do possess is dwarfed by the immensity we simply take from others: as Steven Sloman and Philip Fernbach put it, we never think alone — knowledge lives not in the individual head but in the community, and each of us survives by treating what is accessible in the group as though it were in our own possession, unable in practice to tell the boundary between what we know and what we merely trust others to know. The philosopher John Hardwig drew the conclusion out to its edge: even the scientist, the very model of the rational knower, believes the overwhelming majority of what he holds on the testimony of others whose work he has not checked and could not check; reliance on the authority of others is not a lapse from rational belief but a condition of having beliefs at all in a world this large, so that there are matters on which the rational thing is precisely *not* to make up one's own mind, but to defer to those better placed to know. This is not a weakness of the ignorant. It is the standing condition of every knower without exception, and it is permanent, because the world is too large and a life too short for anyone ever to verify more than a fragment of what he must believe to live.

Put the two halves together and the correction completes itself. The framework can change — the person can genuinely stop trusting the collar. But he cannot stop trusting, because to know anything at all in a world he cannot verify he must take almost all of it from others; and so the act of ceasing to trust one authority is, necessarily and in the same motion, the act of beginning to trust another. The vigilance does not switch off; it re-engages, and reads the cues of the new source. The levers do not release; they re-seat on new materials — a new authority whose marks he now credits, a new community whose belonging he now needs, a new sacred value he will not hear questioned, a new story that now structures the

ground, and, soon enough, a new sunk cost. The verification override does not lift; it re-arms for the new authorised source. The Catholic who leaves the Church and its collar does not thereby step into an unmediated relation to reality; he steps into deference to the credential, the byline, the badge of the movement that woke him, the voice in the feed that finally seemed to be telling the truth — collars all, worn by strangers he cannot verify, certifying authorities he takes on trust. He has changed which unverifiable authority he depends on. He has not escaped depending on unverifiable authority, because there is no escape from it that is compatible with knowing anything, and a mind that had escaped it would know nothing and could act on nothing.

This is the cognitive engine underneath the finding of the paper that precedes this one. *The Altar They Could Not Empty* watched the twentieth century's great regimes try to abolish managed belief by force, and found that the emptied altar always refilled — that the seam through which a people's beliefs are kept cannot be emptied, only recaptured, so that the state which razed the church was compelled to build one, complete with its relics and its saints and its catechism, because the reverence it drove out of the old channel did not evaporate but went looking for a new one and did not have to look far. That paper established the pattern at the scale of nations. This one locates it in the single mind, and shows it is the same pattern, for the same reason. You cannot empty the seat of trusted belief in a person any more than in a people, because a mind with that seat empty would not be a free mind but a mind that knew nothing — it would have refused the testimony on which all its knowledge of the world beyond its own senses depends. The seat cannot be vacated; it can only be occupied by a better tenant or a worse one. And this, finally, is what *The Man Underneath* meant in denying that there is a true self beneath a person's formation waiting to be freed. There is no such self, and there is no such view — no standpoint outside all trusted authority from which a person could at last

see the world plain, unmediated, his own. To revise a formation is never to rise out of trust into the unmediated; it is only ever to move from one trusted set of sources to another, better or worse by the quality and plurality of what one now depends on, but never by escaping dependence, which cannot be done. Seeing through is real. Getting out is not.

PART 7

The Strongest Counter

> *That the strongest case against this paper must now be put at its full strength, and it is a serious one: that the science of trust which this very paper leans on says that people are not gullible, that open vigilance mostly works, that mass persuasion overwhelmingly fails and demagogues mostly preach to their own — so that perhaps there is no "exploit" at all, only the ordinary and largely sound operation of a well-built trust machinery that gets fooled at the margins as any good system does, and that to speak of a permanent vulnerability to elite capture is to resurrect, in new clothes, the very myth of the credulous mass that the science has retired; that this objection is right in everything it asserts and wrong only in what it concludes, because the exploit was never the claim that people are gullible — it is the claim that vigilance is cue-reading, that cues are forgeable, that formation precedes vigilance, and that the whole apparatus is a dependence which can be relocated but never left; and that the honest resting place credits the counter fully — the exploit is not mind-control, backlash is real and constant, and the refilling of the emptied altar is itself the proof that people were never dupes but knowers meeting a real need with the best trusted source on offer*

The strongest objection to this paper comes, fittingly, from the very science the paper is built on, and it must be stated at full strength rather than softened, because a great deal turns on answering it honestly. It runs as follows. You yourself have leaned on Sperber and Mercier to establish that human beings are not credulous — that they run open vigilance, weigh their sources, resist what does not fit,

and are, by default, hard to move. You have granted that mass persuasion mostly fails, that propaganda rarely converts the unwilling, that the demagogue usually preaches to a choir he inherited rather than made. But then how can there be a *permanent vulnerability to elite capture*? If the trust machinery mostly works — if people mostly are not fooled, mostly do resist, mostly see through the crude attempts to move them — then to speak of an "exploit" at all is to smuggle back in, under a new name, the discredited myth of the gullible mass, the passive public helplessly programmed by whoever holds the microphone. The truth, the objection concludes, is far less alarming and far more flattering to the human mind: there is no exploit, only a well-built system of conditional trust that, like any good system, is fooled occasionally at its margins, and the right response to occasional marginal failure is not a theory of permanent capture but ordinary vigilance about the ordinary liars. The previous paper's own evidence can be turned to this end: when the Soviet state threw the entire weight of its apparatus at abolishing religious belief, the belief *survived*, most of the population refusing the official creed at real risk — which looks less like a captured population than like a stubbornly vigilant one that could not be captured at all.

This is a serious objection, and the honest response is to concede everything it asserts and deny only what it concludes. Every factual claim in it is true. People are not gullible; vigilance mostly works; crude mass persuasion mostly fails; backlash against overreach is real and constant; the Soviet believers did refuse, and their refusal is exactly the proof that they were never dupes. The objection goes wrong only in supposing that any of this touches the exploit, and it does not, because the exploit was never the claim the objection refutes. The exploit is not the claim that people are gullible. It has, in fact, spent this entire paper insisting that they are not — that they are vigilant, and that the vigilance is the whole point. The exploit is a different and narrower claim, and each of its four parts survives the

objection intact. First, that vigilance does not verify the world but reads cues that stand in for verification — which the objection does not deny and cannot, since cue-reading is what the science says vigilance *is*. Second, that a cue which can be read can be forged, so that the exploit works by satisfying vigilance with counterfeits rather than by defeating it — which is why the failures that do occur cluster exactly where the cues have been most thoroughly captured, and why the objection's "occasional marginal fooling" is not random but systematically produced wherever an authority controls the signals. Third, that formation precedes vigilance, so that the mature filters the objection celebrates were themselves shaped by an installation that arrived before they existed to screen it — the adult's vigilance is real, but it is *his* vigilance, calibrated by a childhood he did not choose. And fourth, the deepest, that the entire apparatus of vigilance is a dependence, not an independence — a way of choosing whom to trust, never a way of not needing to trust — so that even its every success is the successful placement of trust in some authority the person still cannot verify. The objection proves that people cannot be *stormed*. It does nothing against the claim that they can be *formed* and their cue-environment *captured*, which is the claim the exploit actually makes.

And there is a resting place here that is more honest than either the old paper's grim determinism or the objection's reassurance, and that credits the counter fully. The exploit is not mind-control; it never was, and the previous paper's whole comparative record refutes the mind-control reading as thoroughly as the objection could wish. The believers who kept the faith through seventy years of the League of the Militant Godless were not captured, and the New Atheists who could not argue most of the West out of God were meeting the same limit from the other side: the mind is not a slate to be wiped and rewritten by whoever holds power, and every regime that treated it as one produced backlash, martyrs, and revival rather than the blank population it wanted. The vulnerability is real but it is not

omnipotence, and to state it as omnipotence is to earn exactly the refutation the objection delivers. The precise and defensible statement is this: the exploit is the standing *capturability* of a trust that is necessary, rational, and mostly sound. People must trust; they trust intelligently, on cues; the cues can be forged and the child can be formed; and so the trust that a person cannot do without can be, and across history reliably has been, turned to the advantage of whoever controls the marks and the upbringing — not by making people stupid, which cannot be done, but by meeting their intelligent need for trusted sources with sources that serve the provider. That the emptied altar always refills is not evidence that people are dupes. It is evidence of the opposite: that they were meeting a real and permanent need — the need to hold, on trust, a picture of a world they cannot verify — and that when the authority they trusted was torn away, they did the sane and human thing and found another. The vulnerability and the vigilance are not opposites. They are the same fact seen from two sides. A creature that did not need to trust could not be exploited through trust; but it also could not know anything, and would not be us.

CODA

The Crowded Seam

> *That the conclusion which follows is neither the despair of the old paper nor the complacency of the counter, but the reformist line the whole framework keeps arriving at, now grounded correctly: since no person can be freed from the need to trust authorities he cannot verify — since that need is not a curable ignorance but the permanent price of knowing at scale — the only liberty available is never the abolition of dependence, which is impossible, but the improvement of its structure; that a mind, like a people, cannot empty the seat of trusted belief, but it matters enormously whether that seat is held by a single authority that owns the whole of one's trust and brooks no rival, or by many that compete and check and can be played against one another; and that the work of freedom is therefore to keep the seam crowded — to keep many collars in the room, none able to silence the others, each answerable, each replaceable — so that the dependence which cannot be ended can at least be made contestable, which is the whole of what liberty, honestly understood, has ever been*

What follows from all this is not the despair the earlier paper drifted toward, and not the complacency the counter invites, but the same reformist line the framework keeps arriving at by every road — arrived at now on firmer ground, because the correction shows exactly why the usual hope is idle and exactly what the real work is. The usual hope is that if people could only *see* the mechanism, they would be free of it — that recognition is liberation, that the unmasking of the exploit is the end of it. The correction retires that hope for good. Seeing through one authority does not free you from authority; it delivers you to the next one, because the seeing was done, as all seeing is done, from within some

trust you did not and could not verify. There is no exit from dependence into a clear-eyed autonomy, because the autonomy the hope imagines — the mind that believes only what it has checked for itself — is not a free mind but an empty one, a mind that has refused the testimony on which all knowledge of the wider world depends and therefore knows almost nothing. Freedom cannot mean the end of trusting authorities one cannot verify. It has never meant that, and it never can, and every project that promised it — the Enlightenment fantasy of the man of pure reason, the revolutionary fantasy of the people freed from all managed belief — foundered on the same rock, which is that the seat of trusted belief cannot be emptied, in the person any more than in the nation.

But the seat can be held well or badly, and that difference is the whole of the available liberty, and it is not small. It is one thing for a mind, or a people, to depend on a *single* authority that owns the entire trust and permits no rival — the one church, the one party, the one press, the one voice that has captured every cue at once, so that there is nowhere to stand and nothing to check it against, and the verification override, having no competing authority to trigger a second opinion, simply runs unopposed. It is another thing entirely for the trust to be *distributed* across many authorities that compete, that check and correct and embarrass one another, that can be played off and weighed against each other, none of them able to silence the rest, each of them answerable and replaceable. The dependence is equally total in both cases — the person in the plural world trusts just as much that he cannot verify as the person in the monopolised one — but in the plural world the very thing he cannot escape is made contestable, because the forgery of one authority's cues can be exposed by another's, and the sacred value one source installs can be questioned from the standpoint of a source it does not control, and the story one power tells can be met by a rival telling. This is what the framework has meant, across every paper, by crowding the seam rather than emptying it, and it is the exact application to

the single mind of the reformist conclusion the corpus reaches for the state: not the abolition of the interpreters, which only clears the ground for a worse and lonelier one, but the multiplication of them, so that no single hand owns the whole of what a person is able to believe.

Return, at the end, to the collar. The mistake — the mistake of every project that promised to free people from managed belief — was to think the goal was to get the believer to trust *no* collar, to strip the marks away and leave him standing in an unmediated relation to the truth. That goal is incoherent, because he cannot know his world without taking most of it on trust from marked sources, and a person who trusted no collar would not be enlightened but blind. The goal that is actually available, and actually worth the enormous labour it demands, is to keep *many* collars in the room — many authorities, plural and competing and openly contestable, none able to silence the others, each of them answerable for the trust it asks and replaceable when it forfeits it — so that the deference a person cannot help extending is extended into a field where it can be checked, corrected, and withdrawn, rather than captured whole by one authority that has learned to forge every cue at once. The exploit cannot be closed; it is the underside of the trust that makes us knowers, and it will be with us as long as we are the animal that knows by trusting. What can be built, and defended, and lost, and built again, is the crowded seam — the room full of contending collars — in which the dependence that cannot be ended is at least made to answer. That is not freedom from the exploit. There is no such freedom, for the mind we actually have. It is the only freedom there is: not escape from trust, which is impossible, but the standing, structural contestability of the authorities we cannot do without. Like everything in this work, it is held as tested, not revealed — and this paper, having revised itself by the very trust it describes, is its own first example.

Thematic Bibliography

The framework and its foundations. This paper is a foundational companion to the project's other theory papers and depends on them: *The Machine That Runs Itself* (the floor of force, the robe laid over it to make force cheap, the two dials, the seam, and the reformist conclusion to crowd the seam rather than empty it); *Position Logic* (reading decisions from position, value, and the field of the visible; the sincere, the cynical, and the blinkered); *The Division of Knowing* (the permanence of the seam and of a people's dependence on interpreters, of which the exploit is the cognitive underside); *The Man Underneath* (that there is no true self beneath a person's formation to be freed, and that formation is revisable but never into a view from nowhere); and *The Altar They Could Not Empty* (the twentieth-century attempt to abolish managed belief by force, and the finding that the emptied altar always refills — the pattern this paper locates in the single mind). The framework, the pattern-method, and the term *verification override* are thomas J. Bryan's.

On cultural evolution and obligate social learning. Joseph Henrich, *The Secret of Our Success* (Princeton, 2015), on cumulative culture as the true source of human success and the consequent human reliance on social learning over individual reason; Robert Boyd and Peter Richerson on gene-culture coevolution and the conformist and model-based biases of social learning. Joseph Henrich and Francisco Gil-White, "The Evolution of Prestige," *Evolution and Human Behavior* 22 (2001): 165–196, on freely conferred deference to skilled models and the cues by which we infer whom to copy. On the depth of the copying disposition, the overimitation studies — Victoria Horner and Andrew Whiten (*Animal Cognition*, 2005) and

Derek Lyons and colleagues (*PNAS*, 2007) — showing that human children faithfully reproduce even causally irrelevant actions that other apes skip.

On the vigilance the old account omitted. Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson, "Epistemic Vigilance," *Mind & Language* 25 (2010): 359–393, on the filters by which the mind screens whom and what to trust; Hugo Mercier, *Not Born Yesterday* (Princeton, 2020), on open vigilance and the general failure of mass persuasion — the strongest counter, engaged directly in Part 7; Hugo Mercier and Dan Sperber, *The Enigma of Reason* (Harvard, 2017), on reason as an argumentative and social faculty. Joseph Henrich, "The Evolution of Costly Displays, Cooperation and Religion," *Evolution and Human Behavior* 30 (2009): 244–260, on credibility-enhancing displays — the costly-sincerity cue.

On the five levers. Stanley Milgram, "Behavioral Study of Obedience" (1963) and *Obedience to Authority* (1974), with the essential modern reinterpretation by Stephen Reicher and Alexander Haslam — including Haslam, Reicher, and Millard, "'Happy to have been of service'" (*British Journal of Social Psychology*, 2015) — recasting the results as engaged followership grounded in identification rather than blind obedience, and Gina Perry's *Behind the Shock Machine* (2013) on the methodological problems; Solomon Asch on conformity to the unanimous majority, and its limits. Philip Tetlock, "Thinking the Unthinkable: Sacred Values and Taboo Cognitions," *Trends in Cognitive Sciences* 7 (2003): 320–324, and Scott Atran and Jeremy Ginges on sacred values and the devoted actor (*Science*, 2012) for the sacred-value lock. Jonathan Haidt, "The Emotional Dog and Its Rational Tail" (*Psychological Review*, 2001) and *The Righteous Mind* (2012), on intuition leading and reasoning following, for the narrative and intuitive channel. Daniel Kahneman, *Thinking, Fast and Slow* (2011), for the dual-process vocabulary, used with the caution the ongoing debate over that model warrants.

On the division of cognitive labour — the linchpin of the correction. Leonid Rozenblit and Frank Keil, "The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth," *Cognitive Science* 26 (2002): 521–562; Steven Sloman and Philip Fernbach, *The Knowledge Illusion: Why We Never Think Alone* (2017), and Fernbach and colleagues on the illusion of understanding underwriting political extremism (*Psychological Science*, 2013); John Hardwig, "Epistemic Dependence," *The Journal of Philosophy* 82 (1985): 335–349, and "The Role of Trust in Knowledge" (1991), on the ineliminable dependence of even the rational and the scientific knower on the testimony of others. On the cultural variability of the framework that the architecture underlies, Joseph Henrich, *The WEIRDest People in the World* (2020).

On scale, mismatch, and coalition. Robin Dunbar, "Coevolution of Neocortical Size, Group Size and Language in Humans" (*Behavioral and Brain Sciences*, 1993), for the social-brain hypothesis and the contested ~150 figure; Norman Li, Mark van Vugt, and Stephen Colarelli, "The Evolutionary Mismatch Hypothesis" (*Current Directions in Psychological Science*, 2018); and the coalitional-psychology literature (John Tooby, Leda Cosmides, and Robert Kurzban) on the flexible encoding of group membership.

Modern and copyrighted sources are paraphrased and credited, with quotation kept short; contested findings are flagged as contested. The framework is held, as the whole project holds it, as tested, not revealed — and this revision is its own demonstration, having corrected an earlier formulation not by escaping trust but by re-grounding on evidence its authors now, in turn, depend upon.